Posudek oponenta diplomové práce

Jméno a příjmení:	Bc. Mario Kamburov
Název tématu:	Optimální metody dataminingu pro zpracování semistrukturovaných medicínských dat
Vedoucí práce:	Ing. Petr Včelák (NTIS)

Obsah práce

Obsahem předložené práce byla analýza a návrh vhodné metody pro automatizovanou kontrolu a korekci medicínských dat s cílem nalézt a odstranit překlepy/chyby v textech i nahradit různé typy zkrácených slov za slova v plném znění a se správným významem. Autor vychází z existujících slovníků pro český a latinský jazyk, z vlastní analýzy jemu dostupných textových medicínských dat a několika metod dataminingu, které v práci porovnává.

Logická struktura

Diplomová práce je přehledně strukturována. Seznamuje čtenáře se způsoby tvorby zkratek i s nutností data nejprve předzpracovat, dále se věnuje problematice fulltextového vyhledávání, a výčtu čtyř metod pro datamining. V osmé kapitole popisuje implementaci následovanou porovnáním výsledků jednotlivých metod, závěrečným hodnocením a závěrem.

V některých textech zbytečně zabíhá do oblasti, která s jeho tématem nesouvisí nebo příliš zobecňuje. Např. v kapitole 3.6.1 pojem big data a heterogenita dat v kontextu diplomové práce s čistě textovými daty není objasněna. S řešenou problematikou nesouvisí ani kapitola 3.6.2 Etnické nebo právní problémy, kde navíc ani tentokrát žádný etnický problém uveden není.

Rozsah a formální úroveň

Autorových 74 stran textu odpovídá rozsahu 61 normovaných stran dle metodiky KIV. V textu se objevuje řada gramatických chyb, překlepů i typografické chyby. Vyskytují se také neúplné věty, některé zase postrádají smysl nebo si i protiřečí v rámci jednoho odstavce. Přesto došlo oproti předchozí diplomové práci k výraznému zlepšení formální úrovně dokumentu.

Dokument obsahuje seznam obrázků a tabulek. Příklady a výpisy kódu jsou přímo v textu a nejsou jednoznačně označeny číslem s titulkem, to je pouze u obrázků a tabulek. Současně by pro příklady mohl být použit menší font nebo kompaktnější provedení výpisů, protože místy zabírají více jak polovinu strany. V příloze je ukázkový text před korekcí a po korekci, schéma architektury aplikace a testovací výsledky. Obrázek 12.2. rozhodovacího stromu považuji za zbytečný, zejména z důvodu jeho nečitelnosti. Měl být spíše přiložen jen v elektronické podobě ve větším rozlišení. Na CD jsou přiloženy soubory slovníků (bez informace o licencích), analýza frekvence slov v lékařských datech a také soubory potřebné pro naplnění popisovaných databází včetně dokumentace pro jejich instalaci a zprovoznění. Dostupné jsou na CD také výsledky testování použitých metod dataminingu.

Práce s literaturou

V seznamu literatury je 32 položek, které nejsou vhodně seřazeny dle pořadí výskytu v textu. Uvedené zdroje lze považovat za relevantní. Přesto zejména v teoretické části musím *vytknout nedostatečné citování zdrojů.* Pro řadu tvrzení není uveden žádný zdroj. Za všechny např. kapitola 4.3 (Modely zpracování fulltextových semistrukturovaných dat) neobsahuje relevantní citaci k žádnému z uvedených modelů, vzorců, ani citaci obrázku na str. 26. Podobně také příklad na str. 33, který je převzatý z dokumentace PostgreSQL, ale není uveden žádný zdroj.

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky

Kvalita řešení a dosažených výsledků

Na první pohled je zřejmé, že diplomant zapracoval výtky z předchozího posudku. Realizace spočívá v klient-server aplikaci, kdy klientská část je vlastní plugin v CKEditoru a z uživatelského pohledu fungující jako obdoba známé kontroly pravopisu a gramatiky. Po aktivaci pluginu/funkce je celý text odeslán na server, který vyhodnotí slova a vrátí neznámá slova/zkratky, jež jsou následně v editoru označeny. Je možné použít návrhy oprav, nebo slovník na serveru přímo doplnit o nový záznam. Výsledky všech metod autorovi vychází prakticky shodně, což může být způsobeno malým vzorkem dat k testování i tím, že jde o data převážně z jedné oblasti medicíny, kde se velmi často opakují stejné texty.

Zdrojové kódy jsou přehledně rozděleny do balíků a názvy tříd i metod jsou v anglickém jazyce a srozumitelné. JavaDoc komentáře jsou spíše výjimkou nebo byly pouze automaticky vygenerovány. Názvy souborů pro import dat, řetězce pro připojení k databázím a další nastavení jsou pevně zapsána ve zdrojových souborech.

Opravy nebo návrhy oprav textů jsou automatické, ale vyžadují interaktivní uživatelské prostředí webového prohlížeče s CKEditorem. Řešení v doméně medicínských dat pro reálné nasazení a použití si zcela jistě vyžádá (v první fázi provozu) přímou spolupráci lékařů, aby asistovali u oprav a doplnění slovníku.

Splnění zadání

Zadání a zásady pro vypracování diplomové práce diplomant splnil bez výhrad.

Dotazy k práci

- Jak konkrétně probíhalo uváděné manuální čištění, filtrování a promazávání duplicitních dat v Excelu (vámi zmiňované v 2.1.1)? Jaký rozsah dat byl na počátku a jaký po dokončení?
- V kapitole 5.7 uvádíte testování výkonnosti. Proč jste zvolil pouze 3 pokusy? Za jakých podmínek a jak testy probíhaly (např. restart mezi pokusy, co vše v době testu běželo)?

Závěr

Při hodnocení se rozhoduji mezi stupni velmi dobře a dobře. Dokument, teoretická část s formálními nedostatky a zejména nedostatečné citování zdrojů, je spíše na hodnocení stupněm dobře. S ohledem na praktickou část se celkově přikláním k lepšímu hodnocení.

Navrhuji hodnocení známkou *velmi dobře* a práci *doporučuji k obhajobě*.

V Plzni 23. 8. 2016

Ing. Petr Včelák
NTIS, ZČU

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky